

NIST Open Machine Translation 2014 Evaluation Plan (OpenMT14)

1 INTRODUCTION

The 2014 NIST Open Machine Translation evaluation (OpenMT14) continues the ongoing series of evaluations of human language translation technology. NIST conducts these evaluations in order to support MT research and help advance the state of the art in MT technology. To do this, NIST:

- Defines a set of translation tasks,
- Collaborates with the data providers¹ to provide corpus resources to support research on these tasks,
- Creates and administers formal evaluations of MT technology,
- Provides evaluation utilities to the MT community, and
- Coordinates workshops to discuss MT research findings and results of task performance in the context of these evaluations.

These evaluations provide an important contribution to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of translating between human languages. To this end, the evaluations are designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible to those wishing to participate.

The 2014 evaluation is a pilot evaluation focusing on exploring translation system limitations as well as measurement limitations on informal data genres. Highlights of OpenMT14 include:

- Evaluation on informal data genres SMS/Chat and Conversational Telephone Speech (CTS) for Arabic-to-English and Chinese-to-English,
- Inclusion of audio input track, and
- Explore common MT measurement techniques on these informal data genres.

Participation in the evaluation is invited for all researchers who find the tasks and the evaluation of interest. There is no fee for participation. However, participation in the evaluation requires participation in the follow-up workshop.² All OpenMT14 participants must attend the evaluation workshop and be prepared to discuss their system(s), results, and their research findings in detail. This workshop is restricted to the group of registered OpenMT14 participants, data providers and representatives of supporting government agencies.

To participate in the evaluation, sites must officially register with NIST³ and agree to the terms specified in the registration form. For more information, visit the NIST OpenMT14 website.⁴

2 TRAINING CONDITIONS

MT R&D requires language data resources. System performance and R&D effort are strongly affected by the type and amount of resources used. Therefore, OpenMT14 has two different resource categories as conditions of evaluation. They differ solely by the specification of the data that may be used for system training and development. These evaluation conditions are *Constrained Training* and *Unconstrained Training*, as implemented in previous OpenMT evaluations.

As in previous OpenMT evaluations, some training data are provided by the LDC. All participants are required to sign a license agreement⁵ governing the use of LDC's data resources available for system development in preparation for OpenMT14. Participants must fully comply with all requirements that are (1) stated in this evaluation plan, (2) stated on the registration form, and (3) stated on the LDC license agreement, in order to retain rights to data obtained under the LDC license agreement.

2.1 CONSTRAINED TRAINING

Systems entered in the Constrained Training condition allow for direct comparisons of different algorithmic approaches. System development must adhere to the following restrictions:

Only data listed in the LDC data license agreement may be used for core MT engine development in the constrained training condition. OpenMT14 does not place a language specific restriction on the LDC data resources; that is, a site participating in Arabic to English may use Chinese to English data as long as that data is listed in the LDC data license.

¹ <http://www.ldc.edu> <http://www.sdl.com/research/language-technology>

² There is a registration fee associated with attending the evaluation workshop. This fee does not include travel or accommodation expenses.

³ http://www.nist.gov/itl/iad/mig/upload/OpenMT14_Registration.pdf

⁴ <http://www.nist.gov/itl/iad/mig/openmt14.cfm>

⁵ http://www.nist.gov/itl/iad/mig/upload/OpenMT14_LDCAgreement.pdf

Resources that assist the core engine (such as segmenters, tokenizers, parsers, or taggers) are not subject to the same restriction. If such additional resources are used, they must be listed in the system description.

2.2 UNCONSTRAINED TRAINING

Systems entered in the Unconstrained Training condition may demonstrate the gains achieved by adding data from other sources. This training condition allows for more creativity in system development. System development must adhere to the following restrictions:

Data must be publicly available, at least in principle.⁶ This ensures that research results are broadly applicable and accessible to all participants. Participants must specify in their system descriptions what data they used.

3 DATA SET

The OpenMT14 evaluation data will have approximately the following volume:

Table 1: Data volume for OpenMT14 test sets.

Language Pair	Genre	Volume (words)
Arabic-to-English	SMS/Chat	25,000
	CTS	5,000
Chinese-to-English	SMS/Chat	25,000
	CTS	5,000

Each data set has one gold standard reference. Additionally, HyTER network will be created for approximately 5,000 words for each language-pair and genre.

4 INPUT TRACKS

OpenMT14 will offer two input tracks: audio and text. If the audio track is chosen, participants are required to process both tracks.

4.1 AUDIO SOURCE INPUT

For the audio input track, system will process from the audio of the telephone conversations. Whether segmentation will be given is still under discussion.

4.2 TEXT SOURCE INPUT

For the text input track, system will process the SMS/chat messages as well as the text transcripts of the telephone conversations.

5 PRIMARY AND CONTRASTIVE SUBMISSIONS

OpenMT14 allows participants to submit exactly one primary system submission for each language pair registered. There is no limit (within reason) on the number of contrastive submissions.

At the time of submission exactly one system must be identified as the primary system, for each given language pair. Only primary systems will be compared and contrasted across sites in NIST's reporting of results.

Contrastive systems are encouraged to test significant alternatives to the primary system. NIST discourages contrastive entries that represent mere tweaks and minor parameter setting differences.

6 PERFORMANCE MEASUREMENT

OpenMT14 will use several automatic metrics and, time/resource permitting, will investigate several semi-automatic metrics as well as human assessments of system translations to understand measurement limitations on informal data genres. Unlike previous years, there will be no official primary metric. The results of the scoring techniques will be included as part of the public release of results.

6.1 METRICS

Metrics under consideration are:

- BLEU⁷ – This technique scores a translation according to the N-grams that it shares with one or more reference translations of high quality. In essence, the more co-occurrences, the better the translation. An N-gram, in this context, is simply a *case sensitive* sequence of N tokens. (Words and punctuation are counted as separate tokens.) NIST will compute case-sensitive BLEU scores using NIST's publicly available *mteval* software⁸.

⁶ Data limited to government use, such as the FBIS data, is deemed to be publicly available and admissible for system development.

⁷ Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu (2001). "Bleu: a Method for Automatic Evaluation of Machine Translation". This report may be downloaded from URL <http://domino.watson.ibm.com/library/CyberDig.nsf/home> (keyword RC22176).

⁸ <ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a-20091001.tar.gz>

- METEOR⁹ – This technique scores a translation according to word-to-word matches between the system and reference translations but also includes a set of language specific weights tuned to the target language and the ability to incorporate stemming and synonymy.
- TER¹⁰ – This technique scores a translation according to the number of transformations that must be performed on the system translation such that it has the same word ordering as the reference translation. NIST will compute case-sensitive TER scores using UMD's publicly available tercom software¹¹. NIST may also compute human-targeted version of TER (HTER).
- HyTER¹² – This technique is similar to TER but makes use of large networks of reference translations.
- MEANT¹³ – This technique scores a translation according to the semantic role fillers that match those of one or more reference translations of high quality. In essence, the more semantic role matches, the more useful the translation. The semantic roles answer "who did what to whom, when, where and why". NIST may also compute human version of MEANT (HMEANT).

6.2 HUMAN ASSESSMENTS

OpenMT14 will use a participant-volunteer model for human assessment. If a site wishes for its system output to be included in the assessments, it will be required to perform some assessments of OpenMT14 system translations. NIST will provide an assessment tool for download; judges will return their assessments to NIST. Participation in the human assessments must be indicated on the OpenMT14 registration form. Only primary submissions will be eligible for inclusion in the human assessments.

The human assessments for OpenMT14 will consist of rankings of system outputs at the segment level.

Participants who would also like to participate in the human assessments must submit a separate registration form for human assessments by the registration deadline of April 30, 2014.¹⁴

7 SCHEDULE (TENTATIVE)

- February 1, 2014: Initial evaluation plan available
- February 1, 2014 – June 13, 2014: Registration period (early registration highly encouraged)
- February – June 13, 2014: Training data available from LDC with incremental releases of new genres until the end of May 2014
- June 16 – 20, 2014: Dry run period
- July 14 – 18, 2014: Main evaluation period for audio track; output due at NIST July 18, 11.59am ET
- July 21 – 25, 2014: Main evaluation period for text track; output due at NIST July 25, 11.59am ET
- August 4 – 15, 2014: Human assessment period
- August 28 – 29, 2014: Workshop in the Washington DC area; participants to bring system description as handout
- September 9, 2014: Official public release of results

8 EVALUATION PROCEDURES

The OpenMT14 evaluation process includes a number of mandatory steps; please see the schedule in section 7 for the dates for each of these:

- 1 Register to participate. Each site electing to participate in the evaluation must register with NIST.
- 2 Sign LDC's data license agreement and return it to LDC. **Even if not selecting any training data, participants must sign the agreement to receive the evaluation data, which are listed on the agreement.**
- 3 Receive the dry run source data from NIST. Dry run source data will be sent to evaluation participants at the beginning of the dry run period.
- 4 Perform the dry run translation. Each site must run its translation system(s) on the entire dry run set for each language pair attempted.
- 5 Return the dry run translations to NIST according to the instructions.
- 6 Receive the evaluation source data from NIST. Source data will be sent to evaluation participants at the beginning of the evaluation period. Inspection and manipulation of the evaluation data before the end of the evaluation period are prohibited.
- 7 Perform the translation. Each site must run its translation system(s) on the entire test set for each language pair attempted.

⁹ Michael Denkowski and Alon Lavie, "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems", Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation, 2011.

¹⁰ Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas, 2006.

¹¹ <http://www.cs.umd.edu/~snover/tercom/>

¹² Dreyer, Markus, and Daniel Marcu. "Hyter: Meaning-equivalent semantics for translation evaluation." Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2012.

¹³ Lo, Chi-kiu, and Dekai Wu. "MEANT: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Vol. 1. 2011.

¹⁴ http://www.nist.gov/itl/iad/mig/upload/OpenMT14_HumanAssessmentRegistration.pdf

- 8 Return the translations to NIST according to the instructions.
- 9 Complete assigned human assessments, if participating.
- 10 Receive the evaluation results. NIST will score the submitted system translations and distribute the evaluation results to the participants.
- 11 Submit a system description (see section 12).
- 12 Be ready to prepare an oral or poster presentation for the workshop; NIST will contact selected presenters in ample time before the workshop.
- 13 Attend the evaluation workshop. NIST sponsors a follow-up evaluation workshop where evaluation participants and government sponsors meet to review evaluation results, share knowledge gained, and plan for the next evaluation. A knowledgeable representative from each participating site is required to attend this workshop and be ready to describe their technology, research, and findings. Attendance at the workshop is restricted to evaluation participants and government sponsors of MT research.

Handling of late and debugged submissions: Scores on submissions received at NIST after the submission deadline, as well as submissions that were debugged beyond formatting errors after an initial submission, will not be listed in the official public release of results. The respective sites will be listed in the release as having participated with a late and/or debugged submission. Such submissions will be scored as time permits and may be reported at the evaluation workshop.

9 NIST OPENMT DATA FORMAT

NIST has defined a set of XML tags that are used to format MT source, reference, and translation files for evaluation. Translation systems must be able to input the source documents and output translations that meet these formatting standards. All NIST OpenMT source, reference, and translation files have an *xml* extension; their format is defined by the current XML DTD.¹⁵ NIST requires that all submitted translation files are well-formed and valid against the above-mentioned DTD.

9.1 TEXT SOURCE FILE FORMAT

A source file contains one single *srcset* element, immediately beneath the root *mteval* element. The *srcset* element has the following attributes:

- *setid*: The dataset.
- *srclang*: The source language. One of: Arabic, Chinese.

The *srcset* element contains one or more *doc* elements, which have the following attributes:

- *docid*: The document name.
- *genre*: The data genre. One of: sms, chat, cts_text, cts_audio.

Each *doc* element contains several segments (*seg* elements). Each segment has a single attribute, *id*, which must be enclosed using double quotes.

One or more segments may be encapsulated inside additional elements, such as (but not limited to) *hl*, *p*, or *poster*. Only the native language text that is surrounded by a *seg* start-tag and its corresponding end-tag is to be translated.

Sample source file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.7.dtd">
<mteval>
  <srcset setid="sample_set" srclang="Arabic">
    <doc docid="sample_document_1" genre="sms">
      <seg id="1">ARABIC SENTENCE #1</seg>
      <seg id="2">ARABIC SENTENCE #2</seg>
      ...
    </doc>
    <doc docid="sample_document_2" genre="sms">
      <seg id="1">ARABIC SENTENCE #1</seg>
      ...
    </doc>
    ...
  </srcset>
</mteval>
```

¹⁵ <ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.7.dtd>

9.2 AUDIO SOURCE FILE FORMAT

TBD

9.3 REFERENCE FILE FORMAT

A reference file contains one or more refset elements, immediately beneath the root mteval element. Each refset element contains the following attributes:

- **setid**: The dataset.
- **srclang**: The source language. One of: Arabic, Chinese.
- **trglang**: The target language, English.
- **refid**: The current reference.

Each refset element contains one or more documents, which, in turn, contain segments. The foat of the document elements and their subsequent child elements is exactly the same as described in section 0 above for the source file.

Sample reference file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.7.dtd">
<mteval>
  <refset setid="sample_set" srclang="Arabic" trglang="English" refid="reference01">
    <doc docid="sample_document_1" genre="sms">
      <seg id="1">ENGLISH REFERENCE TRANSLATION #1</seg>
      <seg id="2">ENGLISH REFERENCE TRANSLATION #2</seg>
      ...
    </doc>
    <doc docid="sample_document_2" genre="sms">
      <seg id="1">ENGLISH REFERENCE TRANSLATION #1</seg>
      ...
    </doc>
    ...
  </refset>
  ...
</mteval>
```

9.4 TRANSLATION (TEST) FILE FORMAT

A translation file contains one or more tstset elements, immediately beneath the root mteval element. Each tstset element contains the following attributes:

- **setid**: The dataset.
- **srclang**: The source language. One of: Arabic, Chinese.
- **trglang**: The target language, English.
- **sysid**: A name identifying site and system (see section 10.2 for requirements).

The content of each tstset element is exactly the same as described previously for the source file format and the reference file format.

Sample translation (test) file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.7.dtd">
<mteval>
  <tstset setid="sample_set" srclang="Arabic" trglang="English" sysid="NIST_ara2eng_primary_cn">
    <doc docid="sample_document_1" genre="sms">
      <seg id="1">ENGLISH SYSTEM TRANSLATION #1</seg>
      <seg id="2">ENGLISH SYSTEM TRANSLATION #2</seg>
      ...
    </doc>
    <doc docid="sample_document_2" genre="sms">
      <seg id="1">ENGLISH SYSTEM TRANSLATION #1</seg>
      ...
    </doc>
    ...
  </tstset>
  ...
</mteval>
```

10 SUBMISSION OF RESULTS

Participants may submit output from multiple systems for a given language pair, training condition, and input track. One system must be declared as primary at the time of submission and all other as contrastive.

Each configuration (language pair, training condition, and input track) is considered as a single experiment and is identified by an experiment identifier (EXP-ID). See section 10.1 for the format of the EXP-ID. All experiments are to reside in a single submission file. See section 10.3 **Error! Reference source not found.** for the format of the submission file.

Submission will be made via FTP. If more than one submission is made, the last submission replaces all previous submissions. Submissions that fail validation will be returned to participants for correction. A validation script will be made available in the near future to help participants check their submission prior to sending it to NIST. Late and/or debugged submissions will be documented and scored but will not be compared to other on-time submissions in NIST's reports.

10.1 EXPERIMENT IDENTIFIER (EXP-ID)

The system output files are organized by experiment identifier (EXP-ID) directory. The EXP-ID directory has the following format:

`<year>_<evaltype>_<site>_<langpair>_<train>_<input>_<systype>`

Where:

- `year`: The evaluation name. This year it's `openmt14`
- `evaltype`: The evaluation being performed, referring to `DryRun` or `Main`.
 - One of: `dryrun`, `eval`
- `site`: The unique ID **assigned by NIST** to the site upon registration
- `langpair`: The language pair attempted in this submission
 - One of: `ara2eng`, `chi2eng`
- `train`: The training condition, referring to `Constrained` or `Unconstrained` training
 - One of: `cn`, `un`
- `input`: The input modality, referring to `Audio` or `Text`
 - One of: `audio`, `text`
- `systype`: The type of system of the particular submission. A primary submission must always be present.
 - One of: `primary`, `contrastX` where `X` is a positive integer `1..N`

Example of a well-formed EXP-ID directory: `openmt14_eval_nist_ara2eng_cn_audio_primary`

10.2 FILE NAMING

The system output files must comply with the following naming convention:

`<base>.xml`

Where:

- `base`: The base filename of the test file and should match the base of the input test file.

10.3 SUBMISSION FILE

All EXP-ID directories must reside inside a submission file. The submission file has the following format:

`<year>_<evaltype>_<site>_<subnum>.tgz`

Where:

- `year`, `evaltype`, and `site`: Same as described in section 10.1 above
- `subnum`: The submission number. The initial submission must be `01`. If more than one submission is made, the last submission replaces all previous submissions. Subsequent submissions are to be numbered consecutively (`02`, `03`, etc.).

10.4 SUBMISSION INSTRUCTIONS

The submission for each language pair must be compressed as follows:

- Create an experiment directory for each experiment
 - `mkdir openmt14_eval_nist_ara2eng_cn_audio_primary`
 - `mkdir openmt14_eval_nist_ara2eng_cn_audio_contrast1`
- Place the system output files in the corresponding experiment directory
 - `cp <system output files> openmt14_eval_nist_ara2eng_cn_audio_primary`
 - `cp <system output files> openmt14_eval_nist_ara2eng_cn_audio_contrast1`
- Create a submission directory
 - `mkdir openmt14_eval_nist_01`
- Place all the experiment directories in the submission directory

- o mv openmt14_eval_nist_ara2eng_cn_audio_primary openmt14_eval_nist_01
- o mv openmt14_eval_nist_ara2eng_cn_audio_contrast openmt14_eval_nist_01
- Tar and zip the submission directory
 - o tar zcfv openmt14_eval_nist_01.tgz openmt14_eval_nist_01
- FTP the compressed tar file to jaguar.ncsl.nist.gov/openmt/incoming using anonymous FTP
 - o ftp jaguar.ncsl.nist.gov (anonymous login with email as password)
 - o binary
 - o cd openmt/incoming
 - o put openmt14_eval_nist_01.tgz
 - o bye

11 DRY RUN

TBD

12 SYSTEM DESCRIPTIONS

Participants are required to submit system descriptions of the MT systems used for their submissions. Please use NIST's template¹⁶ for system descriptions. System descriptions should be submitted in text format, and the file name should reflect the site ID.

13 GUIDELINES FOR PUBLICATION OF RESULTS

NIST Multimodal Information Group's MT evaluations follow an open model to promote interchange with the outside world. The rules governing the publication of NIST Open OpenMT14 evaluation results are the same as were used the previous year.

13.1 NIST PUBLICATION OF RESULTS

At the conclusion of the evaluation cycle, NIST will create a report that documents the evaluation. The report will be posted on the NIST web space and will identify the participants and the scores from various metrics achieved for each language pair and training condition. Scores will be reported by language pair for the data subsets described in section 3, both across genres and separately by genre.

Results from the participant-based human assessments may also be posted.

The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

13.2 PARTICIPANTS' REPORTING OF RESULTS IN PUBLICATIONS

Participants must refrain from publishing results and/or releasing statements of performance until the official OpenMT14 results are posted by NIST.

Participants may not compare their results with the results of other participants, such as stating rank ordering or score difference. Participants will be free to publish results for their own system, but participants will not be allowed to name other participants, or cite another site's results without permission from the other site. Publications should point to the NIST report as a reference.¹⁷

All publications must contain the following NIST disclaimer:

NIST serves to coordinate the NIST OpenMT evaluations in order to support machine translation research and to help advance the state-of-the-art in machine translation technologies. NIST OpenMT evaluations are not viewed as a competition, as such results reported by NIST are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U.S. Government.

Linguistic resources used in building systems for OpenMT14 should be referenced in the system description. Corpora should be given a formal citation, like any other information source. LDC corpus references should adopt the following citation format:

Author(s), Year. Catalog Title (Catalog Number). Linguistic Data Consortium, Philadelphia PA.

For example:

Xiaoyi Ma et al, 2005. Arabic News Translation Text Part 1 (LDC2004T17). Linguistic Data Consortium, Philadelphia PA.

¹⁶ http://www.nist.gov/itl/iad/mig/upload/OpenMT14_SysDescTemplate.txt

¹⁷ This restriction exists to ensure that readers concerned with a particular system's performance will see the entire set of participants and tasks attempted by all researchers.